Xu ZW. Cloud-sea computing systems: Towards thousand-fold improvement in performance per watt for the coming Zettabyte era. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 29(2): 177–181 Mar. 2014. DOI 10.1007/s11390-014-1420-2

Cloud-Sea Computing Systems: Towards Thousand-Fold Improvement in Performance per Watt for the Coming Zettabyte Era

Zhi-Wei Xu (徐志伟), *Fellow*, *CCF*, *Member*, *ACM*, *IEEE*

Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

E-mail: zxu@ict.ac.cn

Received February 14, 2014; revised February 17, 2014.

Abstract We are entering a new era of computing, characterized by the need to handle over one zettabyte (10^{21} bytes, or ZB) of data. The world's capacities to sense, transmit, store, and process information need to grow three orders of magnitude, while maintain an energy consumption level similar to that of the year 2010. In other words, we need to produce thousand-fold improvement in performance per watt. To face this challenge, in 2012 the Chinese Academy of Sciences launched a 10-year strategic priority research initiative called the Next Generation Information and Communication Technology initiative (the NICT initiative). A research thrust of the NICT program is the Cloud-Sea Computing Systems project. The main idea is to augment conventional cloud computing by cooperation and integration of the cloud-side systems and the sea-side systems, where the "sea-side" refers to an augmented client side consisting of human facing and physical world facing devices and subsystems. The Cloud-Sea Computing Systems project consists of four research tasks: a new computing model called REST 2.0 which extends the REST (representational state transfer) architectural style of Web computing to cloud-sea computing, a three-tier storage system architecture capable of managing ZB of data, a billion-thread datacenter server with high energy efficiency, and an elastic processor aiming at energy efficiency of one trillion operations per second per watt. This special section contains 12 papers produced by the Cloud-Sea Computing Systems project team, presenting research results relating to sensing and REST 2.0, the elastic processor, the hyperparallel server, and the cloud-sea storage.

Keywords billion-thread server, cloud-sea computing, cloud-sea storage, elastic processor, REST 2.0

1 Zettabyte Era and Challenges

Evidence abounds that we are entering a new era of computing, characterized by the need to handle over one zettabyte (10^{21} bytes, or ZB) of data. This calls for thousand-fold increases in computing capacities while keeping the energy consumption at a level similar to that of year 2010, via innovations in devices, systems, and applications.

The trend to the Zettabyte era seems evident. In both China and the USA, the largest developing country and the largest developed country, an individual Internet service provider already needs to store and handle exabytes (EBs) of data. A recently published industrial study[1] estimates that over 1 ZB of data were generated in 2012 and 40 ZB will be generated by year 2020 worldwide.

In the academic circle, a recent study[2] reveals that the world's capacities to store, communicate, and process information grew exponentially between 1986 and 2007. Table 1 shows such growths in worldwide total capacities plus our conservative extrapolations to year 2030.

Table 1. Worldwide Total Capacities to Store, Communicate, and Compute Digital Information

Year	Storage	Communication	General-Purpose Computing	Special-Purpose Computing
1986	21 PB	59 PB	0.3 PIPS	0.44 PIPS
2007	277 EB	537 EB	6.39 EIPS	189 EIPS
2030	140 ZB	272 ZB	18 ZIPS	2 570 ZIPS

Note: data for 1986 and 2007 are from [2].

Roughly speaking, these numbers represent three eras: the Peta era, the Exa era, and the Zetta era. In 2007, the global information and communication industry already consumed about 2% of worldwide energy. The Zetta era poses an unprecedented challenge: we

Short Paper

Supported by the Strategic Priority Program of the Chinese Academy of Sciences under Grant No. XDA06010401 and the Guangdong Talents Program of China under Grant No. 201001D0104726115.

must innovate to provide thousand-fold increases in performance per watt. A related work is the GreenTouch Consortium, which is formed to increase telecommunication energy efficiency by 1000 times between 2010 and 2020[3].

Table 2 shows the trends of the worldwide capacities per capita, which are obtained by dividing the Table 1 numbers by the world population numbers of respective years. These per-capita numbers represent three eras: the Mega era, the Giga era, and the Tera era.

Table 2. Worldwide Per-Capita Capacities to Store, Communicate, and Compute Information

Year	Storage	Communication	General-Purpose Computing	Special-Purpose Computing
1986	4.3 MB	12 MB	0.06 MIPS	0.09 MIPS
2007	42 GB	81 GB	1 GIPS	29 GIPS
2030	18 TB	34 TB	2 TIPS	321 TIPS

Note: data for 1986 and 2007 are from [2].

These per-capita numbers reveal potential workload requirements, especial new workloads. For instance, previous data communication is mostly to/from human-facing terminal devices, such as desktop PCs, laptops, smart phones, Pads, and TV sets. But according to Table 2, by 2030 each person on the planet will in average communicate 34 TB a year, i.e., 93 GB per day, roughly equivalent to watching 93 high definition films! Clearly, data for human-facing terminals are only part of the story. Data from devices facing the physical world will drive significant new workloads. Computing must not be restricted to the traditional cyberspace (unary computing), but become ternary computing involving the human-cyber-physical universe[4-5]. Related work includes Cyber-Physical systems (CPS) research, crowdsourcing, human computation, and hybrid human-computer database systems[6-8] .

To summarize, in the Zetta era we face three major technology challenges:

• To increase the capacities of storing, communicating, and processing information by three orders of magnitude, while maintaining the level of energy consumption similar to that of year 2010. In other words, we need to improve the performance per watt by 1000 times.

• To anticipate and embrace new workloads, a significant portion of which will come from the human-cyber-physical ternary computing applications.

• To enable transformative innovations in devices, systems, and applications, while without polluting beneficial IT ecosystems. This is akin to how Web computing is added to Internet computing, not replacing the Internet.

2 Cloud-Sea Computing

To address the above challenges, in 2012 the Chinese Academy of Sciences launched a 10-year strategic priority research initiative called the Next Generation Information and Communication Technology initiative (the NICT initiative). A main research thrust of the NICT program is the Cloud-Sea Computing Systems project. In contrast to the familiar "cloud" side, here "sea" refers to the terminal side (the client side), including human-facing and physical world facing subsystems. The project aims to handling ZBs of data while improving performance per watt by 1000 times, by innovations at the overall systems architecture level, the datacenter server and storage system level, and down to the processor chip level.

The project contains four research components: a computing model called REST 2.0 which extends the representational state transfer (REST) architectural style[9] of Web computing to cloud-sea computing, a three-tier storage system architecture capable of managing ZBs of data, a billion-thread datacenter server with high energy efficiency, and an elastic processor aiming at energy efficiency of one trillion operations per second per watt.

2.1 REST 2.0 Cloud-Sea Architecture

The REST architectural style[9] provides a universal architecture for modern Web computing and much of cloud computing systems today. We call this style REST 1.0, and our extension to cloud-sea computing REST 2.0. The REST 2.0 computing model is illustrated in Fig.1.

The idea of sea computing was proposed by Dr. Mian-Heng Jiang of Chinese Academy of Sciences[10]. In the foreseeable future, network computing systems could benefit from a two-side cooperation: that of the cloud side and the sea side. Cloud computing will evolve slowly, with client devices (mainly human-facing) accessing the cloud through REST interfaces, mostly via the HTTP 1.1 protocol and soon the HTTP 2.0 when the later becomes official standard around year 2015. The client devices will continue to run Web browsers or applications with the REST interface.

In cloud-sea computing, the traditional cloud client expands into a sea side space called the *sea zone* (e.g., a home, an office, a factory manufacturing pipeline). There are one or more client devices inside a sea zone, and each device can be human facing or physical world facing. At any time, there is a special device (e.g., a home datacenter, a smart TV set) designated as the *seaport* of a sea zone. The seaport equipment serves three purposes: 1) a gateway interfacing the sea zone to the cloud, 2) a gathering point of information and

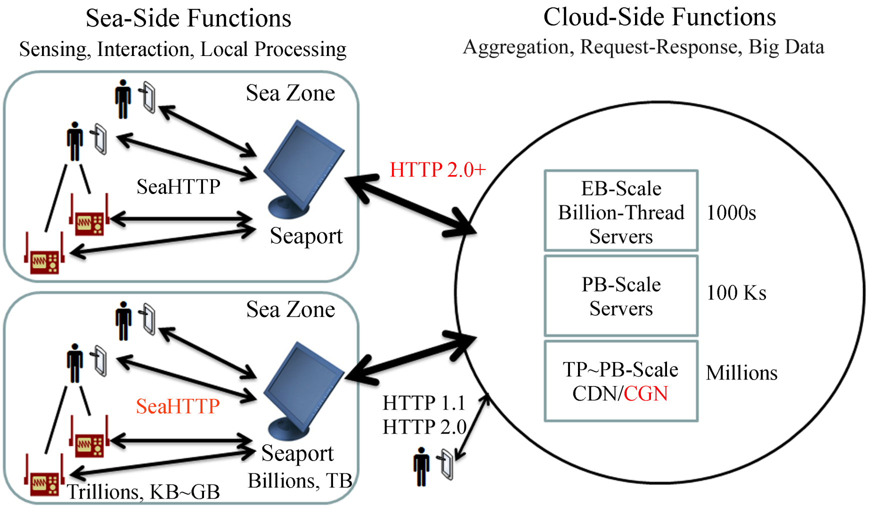

Fig.1. REST 2.0 cloud-sea computing model.

functionalities inside a sea zone, and 3) a shield protecting security and privacy of the sea zone. A device inside a sea zone does not communicate to the cloud directly, but through the seaport, either implicitly or explicitly.

The cloud-sea computing model has four distinct features:

Ternary Computing via Sea Devices. Human and physical world entities interface and collaborate with the cyberspace through the sea side. More specifically, human-cyber-physical collaboration is realized through human facing and physical world facing client devices inside a sea zone. More free forms of usage patterns, such as using a smart phone application to read and control a sensor device in a home, are allowed as long as the entity-sea-cloud pattern is enforced implicitly (hidden from the users), probably through an Internet application service.

Cloud-Sea Cooperation with Locality. A specific network computing system will partition its functions between the sea side and the cloud side. Cloud-side functions include: aggregation of hardware, software, and data resources originally distributed in the global Internet; processing requests coming from sea zones; managing and analyzing big data, etc. Sea-side functions include sensing the physical world, interaction with human and physical entities, and local processing within a sea zone. The cloud-sea computing model encourages locality: hopefully 90% of workloads can be processed within a sea zone, without having to go to the cloud side.

Scalability to ZB and Trillion Devices. The future Internet (or future Web) will consist of many network computing systems offering ternary computing services to billions of people. This future Net will collectively need to support trillions of sea devices and to handle ZBs of data. A potential scenario of such scalability can be realized by the REST 2.0 architecture in the following way:

- The cloud side may have four classes of servers: thousands of high-end servers, each capable of supporting billion-thread parallelism and EBs of data; hundreds of thousands of mid-range servers, each capable of handling PBs of data; millions of content-distribution-network (CDN) servers, each capable of caching TB~PB of data; millions of content-gathering-network (CGN) servers, which play a reverse role of CDN and may provide similar functionality to the middle boxes of software-defined networking.
- The sea side may consist of billions of sea zones. A plausible application case is billions of smart homes where each smart home is a sea zone. A high-end seaport device in a sea zone could manage TBs of data and interface to thousands of sea devices within the sea zone. Thus, collectively we may see billions of sea zones and trillions of sea devices. Human-facing devices (such as a smart phone) can already handle GBs of data. Some physical world facing devices (e.g., a smart watch or other sensor devices) may have more constrained energy and cost requirements, and can only store KBs of data.

Minimal Extension to Existing Ecosystems. The REST 2.0 cloud-sea computing architecture attempts to utilize existing Web computing ecosystems as much as possible. The communication protocol inside a sea zone is a small extension to the HTTP, called seaHTTP. The communication protocol between a sea zone and the cloud is HTTP 2.0, plus some small extensions. The extensions are designed to face the additional challenges brought by human-cyber-physical interactions and constraints in resources, energy, time, space, faults, security and privacy. Such extensions will not violate existing, time-tested REST 1.0 architectural constraints.

2.2 Processor, Server, and Storage Innovations

The cloud-sea computing systems research focuses on three innovative subsystems: elastic processor, hyperparallel server, and cloud-sea storage. Table 3 shows their main performance metrics and innovations, targeting market requirements of year 2020. We will briefly discuss the processor and server research below.

Table 3. Cloud-Sea Computing Systems Innovations

New Subsystem	Capability Metric	Innovative Technology
Elastic processor	Trillion operations per second per watt	FISC architecture, programmable ASIC
Hyper-parallel server	Billion-thread parallelism	Workload processor, load-store interconnect, message interface memory
Cloud-sea storage	Managing ZBs of data	Stable set locality, metadata clusters, storage coding

The main design idea of elastic processor is energy efficient computing by dynamic customization of a microprocessor chip. A new architecture, called Function Instruction Set Computer (FISC), is proposed. Differing from previous RISC and CISC processors, an FISC processor allows the issuing of a "function instruction" (e.g., a machine learning instruction) and execution by a specialized accelerator core. A function could be equivalent to a C function, or even as small grain as a basic block. Mindful of Amdahl's law, the FISC architecture encourages large-grain accelerations but tolerates small-grain acceleration by supporting efficient function instruction issuing within a few clock cycles. At the implementation level, an accelerator is realized as a programmable ASIC. For instance, a machine learning accelerator may be implemented as an ASIC hardware, but can be dynamically programmed to realize different neural network deep learning algorithms at different program execution times.

The hyperparallel server research attempts to improve performance per watt by a hyperparallel server architecture with three characteristics: on-demand customization, billion-thread parallelism, and efficient data movement techniques. The architecture reduces system complexity of conventional servers, and enables software to customize the hardware according to workload needs. This on-demand customization allows application workloads to execute on a befitting bare metal. By judiciously exploring chip-level, node-level, rack-level, and system-level parallelisms, the architecture is scalable to support billion-thread parallelism, which is needed by the workloads of 2020[11]. The architecture provides efficient memory and interconnect techniques to reduce data movement cost.

3 Scanning the Special Section

This special section contains four papers relating to sensing and REST 2.0, three papers on the elastic processor design, three papers on server research, and two papers on storage.

In the article "A Functional Sensing Model and a Case Study in Household Electricity Usage Sensing", Liu *et al.* propose a novel sensing model called SDR (Sampling-Design-Reconstruction). This model introduces the concept of design space to decouple a sensing procedure into two independent modules: the sampling protocol and the reconstruction algorithm, thus improving reusability and upgradeability. The subsequent case study shows that most existing types of household electricity usage sensing systems can be successfully decoupled with their model.

In the article "EasiSMP: A Resource-Oriented Programming Framework Supporting Runtime Propagation of RESTful Resources", Qiu *et al.* devise a novel macro-programming framework utilizing and integrating sensor, mobile tablet, and cloud resources. Experiments of six practical application scenarios show that EasiSMP can reduce lines of code by 85% over the device-based NesC/TinyOS programming method.

In the article "SeaHttp: A Resource-Oriented Protocol to Extend REST Style for Web of Things", Hou *et al.* extend the REST style to future Internet applications which inevitably will contain large numbers of devices. SeaHttp focuses on the parallel processing of resource requests involving multiple devices. Experimental results show that using SeaHttp can reduce energy consumption compared with using the Constrained Application Protocol (CoAP).

In the article " A Task Execution Framework for Cloud-Assisted Sensor Networks", Shi *et al.* suggest building a framework for task execution with the help of the cloud. The framework selects devices and generates an optimized task schedule according to the fuzzy task description by the applications. Experiments show that such a framework significantly reduces sampling cost and energy consumption.

In the article "An Elastic Architecture Adaptable to Various Application Scenarios", Wu *et al.* propose the concept of elastic architecture which can achieve high elasticity (the ratio of the best-case to the worst-case execution times in the processor parameter design space). Performance evaluation on a prototype implementation shows an average improvement of 31.14% on energy-delay product, over a fixed architecture.

In the article "A General-Purpose Many-Accelerator Architecture Based on Dataflow Graph Clustering of Applications", Chen *et al.* propose a novel instruction set architecture called FISC to integrate multiple on-chip accelerators with a general-purpose processor core. Such accelerators are designed to accelerate critical instruction blocks that are most representative according to the dataflow graph clustering of application benchmarks.

In the article "Prevention from Soft Errors via Architecture Elasticity", Yin *et al.* propose a mechanism to proactively prevent soft errors. The key idea is to dynamically adjust processor architectural parameters under the guide of a predictive model establishing a relationship among processor architecture, program execution characteristics, and occurrence of soft errors.

In the article "MIMS: Towards a Message Interface Based Memory System", Chen *et al.* suggest replacing traditional synchronous bus interface of DRAM memory system (e.g., DDR3) with an asynchronous message interface called MIMS, which decouples memory access from memory organization. In addition to ordinary memory requests, other semantic information (such as memory access granularity) can be integrated into packets to achieve more intelligent memory systems. Experimental results on a 16-core simulator validate the advantages of MIMS on performance, energy delay product, and bandwidth utilization.

In the article "Reinventing Memory System Design for Many-Accelerator Architecture", Wang *et al.* propose a wide-I/O-DRAM-based main memory architecture to exploit the sub-rank parallelism in DRAM to support multi-granularity concurrency in the memory access streams issued by abundant on-chip accelerators. This architecture enables an adaptive data fetching mode for accelerators and actively batches their memory requests into a read burst that is serviced in parallel by the DRAM module with a pre-optimized data layout, significantly improving the energy-efficiency of the main memory subsystem.

In the article "A High-Performance and Cost-Efficient Interconnection Network for High-Density Servers", Bao *et al.* present a novel intra-server interconnection network, with associated partition schemes and routing mechanisms. The proposed Hoffman-Singleton network can attain competitive performance as the fully-connected network and cost close to the Torus network.

In the article "SAC: Exploiting Stable Set Model to Enhance CacheFiles", Liu *et al.* improve CacheFiles by utilizing stable sets, a recently observed locality phenomenon. An enhanced CacheFiles system is designed on the pNFS platform. Benchmark results show an order of magnitude I/O performance improvement over existing CacheFiles.

In the article "A Non-Forced-Write Atomic Commit Protocol for Cluster File Systems", Shao *et al.* present an asynchronous atomic commit protocol, named DualLog, to maintain the consistency of distributed metadata update operations in a cluster file system. The protocol improves performance for distributed metadata update operations, as two metadata servers mutually record redundant redo logs, and each update operation incurs only two exchange messages and no sync to disk.

References

[1] Gantz J, Reinsel D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. December 2012. http://idcdocserv.com/1414, Feb. 2014.

[2] Hilbert M, López P. The world's technological capacity to store, communicate, and compute information. *Science*, 2011, 332(6025): 60-65.

[3] The GreenTouch Consortium, GreenTouch Green Meter research study: Reducing the net energy consumption in communications networks by up to 90% by 2020, June 2013. http://www.greentouch.org/index.php?page=green-meter-research, Feb. 2014.

[4] Xu Z, Li G. Computing for the masses. *Communications of the ACM*, 2011, 54(10): 129-137.

[5] Xu Z, Tu D. Three new concepts of future computer science. *Journal of Computer Science and Technology*, 2011, 26(4): 616-624.

[6] Derler P, Lee E A, Vincentelli A S. Modeling cyberphysical systems. *Proceedings of the IEEE*, 2012, 100(1): 13-28.

[7] von Ahn L, Maurer B, McMillen C, Abraham D, Blum M. reCAPTCHA: Human-based character recognition via Web security measures. *Science*, 2008, 321(5895): 1465-1468.

[8] Trushkowsky B, Kraska T, Franklin M J, Sarkar P. Crowd-sourced enumeration queries. In *Proc. Int. Conf. Data Engineering*, April 2013, pp.673-684.

[9] Fielding R T, Taylor R N. Principled design of the modern Web architecture. *ACM Trans. Internet Technol*, 2002, 2(2):115-150.

[10] Jiang M. Urbanization meets informatization: A great opportunity for China's development. *Informatization Construction*, 2010, 6: 8-9. (In Chinese)

[11] Xu Z. How much power is needed for a billion-thread high-throughput server? *Frontiers of Computer Science*, 2012, 6(4): 339-346.

Zhi-Wei Xu received his Ph.D. degree from the University of Southern California, USA. He is a professor of the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. His research areas include high-performance computer architecture and network computing science.